ICCAE 2010 SINGAPORE

Automated Scientific Document Retrieval

University of La Rochelle - Open University Malaysia

JASPAL KAUR

Outline

- Introduction
- Problem Statement
- Research Objectives
- Methodology
- Architecture
- Conclusion

Introduction

- Digital Library
- Information Retrieval
- Scientific Document
- Multimodal Documents

Digital Library

- A distributed and / or online, searchable, organized collection of scientific material.
- Enables searching for electronic content
- Search results are usually shown as a list of items on page.
- Unfortunately, digital library search engine algorithms do not look for information within images / figures.

Information Retrieval

- Users can either locate the information precisely or search for it.
- Searching is either direct or indirect, depending on whether users are familiar with what they want or are just browsing.
- Information retrieval systems or search engines are based on models of retrieval algorithms.
- Text documents are easier to index by their contents rather than multimedia documents.

Scientific Document

- Articles written by experts / scientists in their field are published in journals and other publications.
- These articles can be in printed or electronic form.
- Scientific Literature always include figures for illustration.
- Structure of a document allows for specific indexing methods.
- Contextual queries are possible when document is structured using XML.

Multimodal Documents

- A document conveys information using multiple modalities, including text, images and layout / styles.
- Journal article usually has title, subtitles.
- Indexing and retrieval using only text is the traditional way or Information Retrieval (IR).
- Increasingly important to develop IR techniques for intelligent indexing and retrieval of multimodal documents.

Problem Statement

Personal Digital Library

- Information / documents downloaded from the WWW is kept in folders, named arbitrarily.
- Searchers for this data is time consuming, queries inadequate therefore many useful documents are not referenced again.

Current Systems

- While development of storage devices intensify, intelligent search engines are not.

Research Objectives

The work will pursue the following specific objectives:

- To establish an infrastructure for converting unstructured data into structured documents for indexing.
- To understand the behavior of probabilistic retrieval models with respect to complexity of text document indexing; and to refine and use the models accordingly.
- To develop an online fully integrated system for retrieval of scientific documents.
- To evaluate the above system using known document retrieval assessments and metrics.

Methodology

- Developing intelligent classification algorithms that will be able to unobtrusively elicit user feedback, combine it with contextual and historical evidence, and produce effective structured annotation of new data.
- Summarizing content from a number of textual and semi structured sources, including databases and web pages, in the right way (language, format, size, time) for a specific user, given a task and the user profile.

Multimedia Retrieval Models

- Document Processing
- Data Indexing and Categorization
- Query Processing
- Vector Space Model
- Probabilistic Model
- Content Based Image Retrieval (CBIR)

Document Processing

- In this step, the document is broken into a recognizable, desired retrieval unit.
- Document retrieval could be in the form of title, abstract, authors, summary, references or even paragraphs, if the indexable unit is determined to be a paragraph.

Data Indexing and Categorization (1)

- Potentially indexable elements in the document are identified.
- At this stage, the system requires a set of rules to be executed which control what actions are taken by the algorithm which recognizes 'indexable terms'.
- Indexing of documents should be a continuous operation, so that updates are performed continuously, without building or merging a new index.
- A hash table of words (inverted index) is maintained where an entry contains a compressed version of a word and a pointer to a block.
- A unique identification number for each document in the collection where the term occurs creates a link to each of these documents.
- Then, weights for each term as determined by the IR model being implemented are calculated.

Data Indexing and Categorization (2)

- To determine whether or not a document is pertinent to a particular retrieval process, information must be examined in context.
- Natural language computer interfaces allow users to access complex systems intuitively. Information categorization is the process by which documents are classified into different categories.
- Current technologies use Machine Learning, which uses an inductive process that learns the characteristics of a category from a set of pre-categorized documents .

Query Processing

- The questions posed by the users are represented as queries to the system. Opposed to document processing which occur as a background process, queries occur in real-time, as the user waits for the documents requested. Phrase recognition, insertion of logical operators between terms and expanding the query to include variant terms that refer or relate to the same concept is some of the processes carried out in this step.
- Once the query representation is produced, the matching of documents to the query is carried out. Each document that contains any of the query terms is retrieved. A list of perceived relevant results is shown to the user, where the query is modified based on user-relevance feedback.

Vector Space Model

- The Vector Space Model is the most commonly used model in Document Retrieval System today, due to its consistent, proven performance across multiple implementations on many collections.
- In a vector space model, a document is represented by a vector of terms, and these vectors exist in term space, which is the size of all the unique terms in the collection.
- Each term represents a dimension in this term space and the similarity between a query and a document is measured by the closeness of the query vector and the document vector.
- Term Frequency (TF) / Inverse Document Frequency (IDF) is a weighting scheme that determines that the best indexing terms are those that occur with high frequency in a document (TF) relative to their occurrence in other documents in the collection (IDF).
- Based on this similarity score, the model produces a ranked list of documents in terms of predicted relevance to the query.

Probabilistic Model

- The Probabilistic Model introduced by Robertson and Sparck Jones is based on the assumption that terms occur independently of each other in the documents.
- This model assigns the odds of relevance for each term in a document based on that term's frequency in a set of known relevant documents, and is able to rank documents in terms of their probability of relevance.

Content Based Image Retrieval (CBIR)

- Current CBIR systems generally use primitive features such as color, texture, or logical features such as object and their relationships to represent images.
- This model has been developed for more than a decade and its goal is to search a given image collection for a set of relevant images that are similar to one or more query images.
- Typical CBIR are also built on a vector space model that represents an image as a set of features. The difference between two images is measured through a similarity function between their feature vectors.

Architecture of PDL

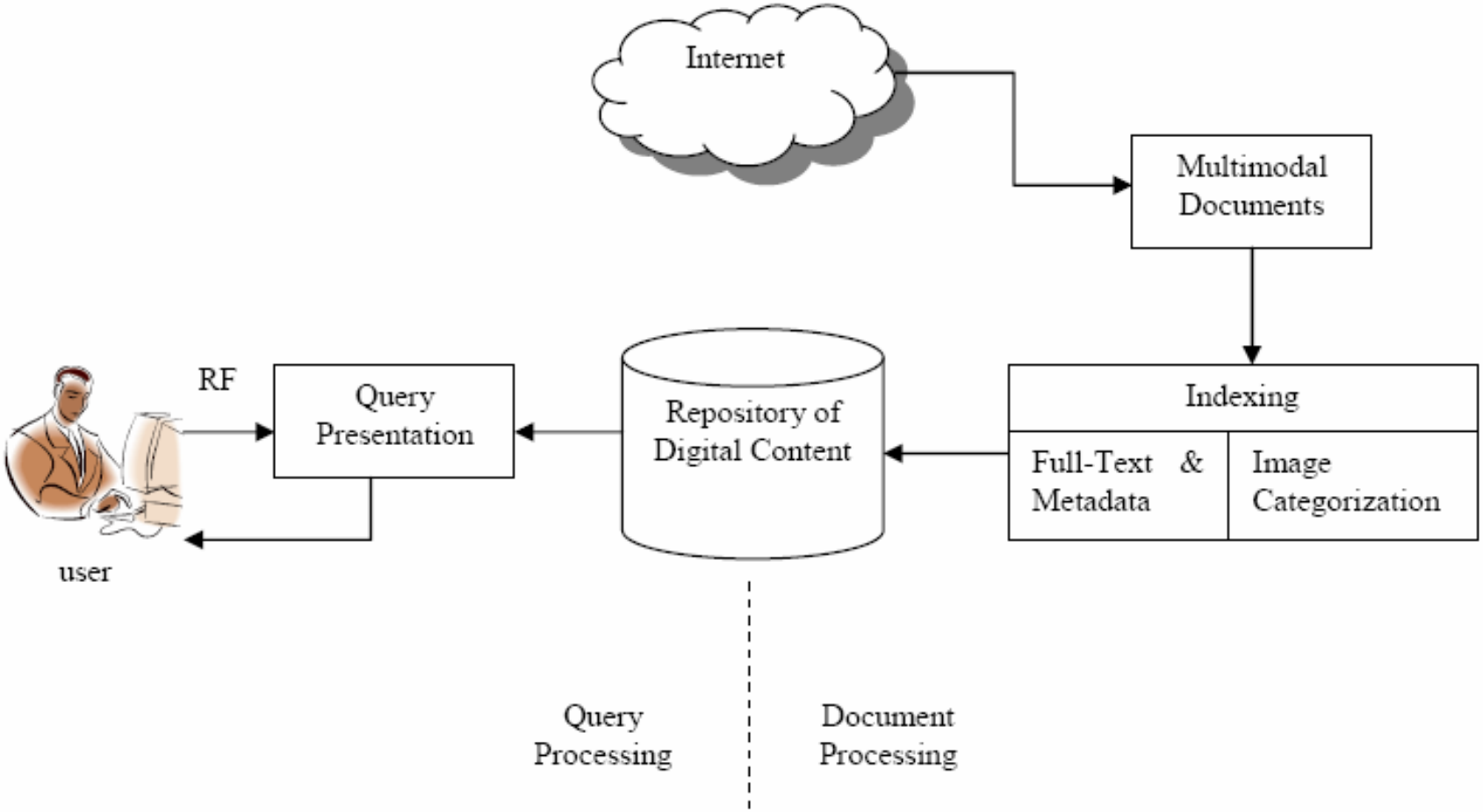

Figure 1. Multimedia-Based Document Retrieval Model

Conclusion

- Greater demand for efficient and effective means for organizing and indexing data is sought so that useful information is retrieved when needed.
- Wide variety of forms in which information is stored and retrieved is a challenge.
- Indexing and searching for multimedia data should be as easy as how text is now.

THANK YOU